2009

Everybody Likes Likert: Using a Variable-Interval Slider to Collect Interval-Level Individual Options

D. Alan Ladd
Washington State University, darin_ladd@wsu.edu

EVERYBODY LIKES LIKERT: USING A VARIABLE-INTERVAL SLIDER TO COLLECT INTERVAL-LEVEL INDIVIDUAL OPINIONS

Research-in-Progress

D. Alan Ladd
Department of Information Systems
Washington State University
Pullman, WA
darin_ladd@wsu.edu

Abstract

As computers become more pervasive in opinion-based surveys, research is required to update existing survey methodologies with current computer capabilities and to begin extending current survey methods by validating additional computer-enabled functionality. This need is particularly acute in the measurement of constructs representative of individuals, such as personality, since current methods were not developed for analysis of individuals. This study addresses the current gap in theoretical justification for measurement of individuals, and then contributes to the development of new functionality to account for this gap. First, it uses computer simulation to explore the overall impact of two types of errors introduced by the number of scale anchors. Second, it proposes the functionality of a new data collection tool called the "variable-interval slider (VIS)," a tool that allows the researcher to account for these two types of errors.

Keywords: Survey methodology, web survey, Likert, continuous scale, variable-interval slider, VIS

Acknowledgement
The author acknowledges the assistance of Monte Schaffer and Anna McNab in clarifying the concepts of interest and specifying the model contained herein. Any errors contained are, of course, fully attributable to the author.

Disclaimer
"The views expressed in this article are those of the author and do not reflect the official policy or position of the United States Air Force, Department of Defense, or the U.S. Government."

EVERYBODY LIKES LIKERT: USING A VARIABLE-INTERVAL SLIDER TO COLLECT INTERVAL-LEVEL INDIVIDUAL OPINIONS

Introduction

It is hard to think of a tool in the modern researcher's belt that is as pervasive, and important, as the computer. Computers support sophisticated data collection, manipulation, analysis, and results reporting in ways that few researchers seventy years ago could have dreamed of. Yet, older scaling techniques such as those based on scales such as Thurstone's (Thurstone and Chave 1929) method of equal-appearing intervals and Likert's (Murphy and Likert 1938) method of summed ratings continue to dominate in spite of their inherent shortcomings. For example:

- Thurstone/Likert scales treat data presumed to exist on a continuum as discrete during data collection, and then re-transform discrete collected data into continuous data for interpretation.
- In some cases these techniques fail to validate assumptions, such equivalent scale values, that some consider essential to interpretation of survey responses (Lodge and Tursky 1979).
- Thurstone/Likert scales assume an individual's ability to accurately estimate an attitude, admittedly representing a region, with a point (Murphy and Likert 1938), possibly introducing an overly precise estimate.
- The validation of these techniques was accomplished for the purpose of sampling populations using between-subjects designs and correcting for error by using the "Law of Large Numbers" (Bock, Velleman, and DeVeaux 2007), possibly creating additional confounds for researchers using these tools to explore individual phenomena such as personality.

As computers become more pervasive in the creation and administration of surveys, research is required to update these existing survey methodologies with current computer capabilities and to begin extending current survey creation methods by validating additional functionality made possible through the use of computers (Dillman 2000). One technique that shows promise is a continuous, sliding scale called a "slider." Though desirable properties of this technique were documented prior to the advent of computer-based coding (Lampert 1979), the prevailing idea amongst survey researchers is that it would be too cumbersome to use, and too difficult to properly code (Russell and Bobko 1992). Recently, however, computer-based survey collection tools have begun to mature to the point that it seems appropriate to re-address the potential of the slider as a viable alternative to forced-choice tools. Beneficial as it may be, use of a computer-based slider presents several unique challenges to the interface designer.

This study explores two error types endemic to the number of scale anchors. First, it discusses two types of error introduced by forced-choice and continuous scales. Second, the results of a computer simulation are reported, showing how the two error types might react when scale anchors are increased. Third, a tool is described that might control or account for the two error types described. Fourth, a study is proposed to use this tool to validate simulation results. The study concludes with remarks about the limitations of the research, as well as implications of the findings to computer-based survey researchers.

Problems with Survey Instruments

The measurement of attitudes with self-reflective instruments is fraught with difficulty, beginning with the nature of attitudes as a latent construct, and ending with the ability of a specific instrument to accurately convey a specific attitude or opinion as it exists in the mind of an individual.[1] Assuming a researcher is able to theoretically justify

[1] This study does not discuss the definition of attitudes or opinions, although it is important to note that there is considerable debate over whether latent attitudes, consisting of both direction and magnitude (Alwin 1992), are transient or persistent (Kirkpatrick 1936; Sherman 1932), and whether or not they are measurable by opinions or behaviors (Thurstone 1928; Thurstone and Chave 1929; House 1934). Instead, term "opinion" is used to sidestep this debate with the understanding that the results might apply equally well to both attitudes and opinions.

opinion measurement with a survey instrument, the question essentially becomes one of measurement error (Dillman and Smyth 2007), and its effect on statistical conclusion and construct validity (Shadish, Cook, and Campbell 2002). This study attempts to describe and measure two error types, labeled A and B for convenience:

- Type A Error: is the scale too coarse?
- Type B Error: is the scale too fine?

Type A Error: Is the Scale Too Coarse?

The ability of an individual to discriminate between items may be quite acute. For example, Cox (1980) reports on multiple studies exploring individual discriminatory ability, the most striking of which is the ability to discriminate between 350,000 different sound frequencies when asked to compare pairs. If a scale has too few choices, an individual is forced to choose (Lehmann and Hulbert 1972; Stroud et al. 1956) between choices that do not accurately represent the individual's "true" opinion, presumed to be continuous and normally distributed over stimuli (Thurstone 1959). Forcing *categorical* judgments versus *continuous*, *quantitative* judgments may result in the loss of data (Lodge and Tursky 1979). This loss may result in data that are *nominal* or *ordinal*, as opposed to the desired *interval* or *ratio* (as defined by Stevens 1946), thereby limiting the statistical tools available for data interpretation (Bock et al. 2007; Dillon, Madden, and Mulani 1983; Lodge, Cross, Tursky, and Tanenhaus 1975; Townsend and Ashby 1984). Rosenthal and Rosnow (2008) criticize the 5-point Likert-type instrument as often misused because of a lack of attention to this concern. The loss of data due to insufficient scale items is termed "Type A Error" in this study.

In general, both Thurstone-type (Thurstone, 1928; Thurstone and Chave 1929) and Likert-type (Murphy and Likert 1939) scales exhibit Type A error. It is worthwhile to note, however, neither Thurstone nor Likert scales were created for use at the individual or single-question level; instead, they were intended for use with large samples, averaging across individuals (between-subjects), where any measurement error introduced by forcing more or less continuous attitudes into ordinal categories would be taken into account by the "Law of Large Numbers." This is important to note because early studies found that reliability and validity was easily established in large-sample between-subjects designs with as few as three scale anchors (Jacoby and Matell 1971), although it is important to note that this assertion remains actively contested (Lehmann and Hulbert 1972). Type A error is particularly worrisome when collecting data on individual opinions, because it is impractical to use the "Law of Large Numbers" to control for this error; in other words, any error created is more difficult to reduce by averaging multiple respondents' answers. Therefore, Type A error threatens two types of validity: 1) construct validity—possibly confusing constructs with the levels of constructs; and 2) statistical conclusion validity—because measures become unreliable, and possibly violate the assumptions of the statistical tests used to analyze the data (Shadish et al. 2002).

Type A error is relatively well-documented. Numerous studies show that increasing the number of scale anchors at addresses many of the underlying internal validity concerns of the Likert-type instrument by increasing the fidelity of the responses to the point that, presumably, an average respondent might not be able to discriminate a finer scale (Lehmann and Hulbert 1972). Monte Carlo simulation showed that the ability of scales to reproduce an underlying distribution leveled off at about 11 items (Mathieson and Doane 2003), although this analysis assumed there was an existing distribution to be modeled, i.e., population values were sampled. The logical conclusion of this stream of research, enabled by modern computer data-gathering techniques, is that a continuous scale would reduce or eliminate Type A Error because continuous scales minimize interval size.

Historically, continuous scale operationalizations used pen and paper techniques. For example, a respondent marked an opinion on an anchored horizontal continuum with a vertical line, or with an "x," intersecting the continuum (Albaum, Best, and Hawkins 1981). The researcher measured the distance of this mark from the origin anchor with a ruler, recording this value. Previous results showed promise: use of fine-grained or continuous scales increased power consistently, though not significantly, higher than that of Likert-type scales (Mathieson and Doane 2003). Further, some argued that the existence of more response choices might reduce some of the positive bias found in Likert-type responses, as well as decreasing kurtosis (Dawes 2008). However, others reported that non-computerized continuous scales were "cumbersome and labor intensive" to use and code (Russell and Bobko 1992). Considering these benefits, it seems possible that a computer-based continuous scale may provide a better solution to measurement error, while minimizing shortcomings of the paper-based approach. Indeed, scales like these are readily available through commercial sites such as Qualtrics; however, computer-based slider functionality is not yet validated, although comparisons abound (see, for example, van Schaik and Ling 2003; van Schaik and Ling 2007).

Type B Error: Is the Scale Too Fine?

Assuming a scale has enough fidelity, and assuming a subject is able to fully articulate his or her opinion (no small concern, as it turns out), then collecting continuous data seems justified. If, however, a scale has too many choices, then an individual's responses may begin to reflect a *region* instead of a *point* on a continuous scale. This "region of indifference" may vary between respondents, may vary from item-to-item on a survey in a given population, may be related to the type of instrument used, and may not remain constant over time. Indeed, in extreme cases, a region of indifference might encompass the entire width of the scale—regardless of the number of choices presented. The creation of artificial precision due to excessive scale items is termed "Type B Error" in this study, represented by a region of indifference on a continuous scale.

A region of indifference creates concerns about the type of data collected. For example, if an individual marks the continuous scale in the general region of "agreement" or "disagreement," i.e., a *nominal-categorical* answer, but the tool records a result that has many more significant digits, e.g., an *interval* response of 3.25 on a scale of 1.00 to 1,000.00, then the data type is incorrectly specified. For this reason, some argue that any given scale might have a natural number of anchors (Green and Rao 1970) in a given population, with as few as 6 anchors accounting for as much as 95% of between-subjects variance. Unfortunately, this does not address the inherently individual nature of this natural number of anchors—a number that could conceivably vary widely within a given population.

Type B error is worrisome when aggregating data on individual opinions because of its ability to change the type of data collected. For example, if the region of indifference is quite large in a given population, this could result in incorrect power calculations. Type B is even more worrisome where violations of data type assumptions might completely prevent reliable measurement. Therefore, Type B error threatens statistical conclusion validity—as measures may be unreliable in measuring the constructs of interest (Shadish et al. 2002).

Summary of Concerns

Table 1 outlines the main points of concern. First, none of instruments currently in use were originally intended for analysis of individuals. Second, only a continuous scale is truly appropriate to measure a hypothesized continuous construct without introducing error due to the scale used (Type A error). Third, though all scale types include Type B error, its visibility and overall impact may be increased when continuous scales are used.

Table 1. Comparison of Opinion-Based Survey Instruments

	Thurstone/Likert	**Continuous**
Original Level of Analysis	Aggregate	Aggregate
Possible Error Concerns	Type A, Type B	Type B
Data Type	Categorical/Ordinal	Quantitative/Interval

Study #1: Simulation

This study modeled and simulated Type A and Type B error, and was guided by the following research question:

- **Research Question (RQ) #1:** How does error due to the number of scale anchors (Type A and Type B error) vary as the number of scale anchors is increased from 2 (end-points only) to ∞ (continuous)?[2]

[2] This study extends studies by Bartholomew and Schuessler (1991), Dunn (1993), and Cox (1980), all of whom discussed Type A error. It extended these three studies in that: 1) it used Cox's ideas to explore the effects of interest at the level of the individual, requiring specification of a different model than used by both Bartholomew/Schuessler and Dunn, and 2) it assumed that the individual's opinion was represented by a region of indifference, instead of a point estimate. In this way, it was possible to build a model that encompassed more theoretical possibilities at the individual level, and best modeled the effects of those possibilities.

Simulation

This study consisted of Monte Carlo simulations developed to model the distribution of average Type A and Type B errors as they interacted with the number of scale anchors as they increased from 2 (dichotomous) to ∞ (continuous) (following Lehmann and Hulbert, 1972). Equation 1 describes Type A error, and Equation 2 described Type B error:

- $$Y_a = \frac{\sum \varepsilon_a}{n}, 0 \leq \varepsilon_a \leq \frac{1}{2X-2} \quad (1)$$
  - Y_a is the average Type A error in terms of the total scale width of 1.0 (range varies from .5 when scale is dichotomous to 0 when the scale is continuous)
  - ε_a is a *uniformly distributed* random variable falling between the upper and lower limits[3]
  - X is the (integer) number of scale anchors, ranging from X = 2 (end-points only) to X = ∞ (continuous)
- $$Y_b = \frac{\sum \left[\varepsilon_b - \left(\frac{1}{X-1}\right)\right]}{n}, 0 \leq \varepsilon_b \leq 1 \quad (2)$$
  - Y_b is the average Type B error, defined as the width between scale anchors subtracted from the region of indifference width, where both are measured in terms of total scale width of 1.0 (range is -1 to 1)
  - ε_b is a *normally distributed* random variable falling between the lower and upper limits[4]
  - X is the (integer) number of scale anchors, ranging from X = 2 (end-points only) to X = ∞ (continuous)

In this way, the study modeled the effect on the two types of error for an individual answering multiple questions over the different scale values. Following the literature, X was held constant at values of 2 through 20, 50, 150, and 100,000 (to approximate ∞). Next, 5,000 questions, representing the full scale of possible attitudes, were administered (converged at 2 significant digits) to assess the overall error. Finally, an average error was calculated over the 5,000 trials. The total number of simulation sets was 22 per error type, and the total number of data points was 220,000. The simulations were generated using Microsoft Excel 2007.

It is important to note that Type B error, though measured in terms of the number of scale anchors, represents a different concept than Type A error.[5] While Type A error is fairly easy to understand in that it is the linear difference between two points on a line, Type B error is represented by a region of indifference that is centered on a point—though for the sake of simplicity the width of this region is then compared to the width of a scale item. An important logical consequence of this convenience is that, even if a region of indifference remains constant, as the number of scale anchors increases the Type B error will also increase. So, Type B error is sensitive to two inputs: 1) the individual respondent's reaction to a question, i.e., the size of the region of indifference, and 2) the number of scale anchors. So, an individual respondent's reaction might encompass the entire scale regardless of the number of anchor points offered, but as the number of anchor points increase, this error would become more *noticeable*.

Data Analysis and Results

Figure 1 shows the simulated relationship between the number of scale anchors and scale error type. The two error types are shown on one graph for convenience of comparison by number of scale anchors. Average Type A error exhibits an inverse-square decay as scale anchors are increased, and average Type B error exhibits a logarithmic increase as scale anchors are increased. These results confirm previous research in that they show average Type A error tapering off to about 3% of overall scale length when 9 anchors are included on a scale. Average Type B error exhibits some notable traits: 1) it is not noticeable until 4 anchors are included on a scale, at which point it covers nearly 17% of overall scale length, 2) it reaches 37% at 9 anchors, and 3) it levels off at about 48% at 50 anchors.

---

[3] Assumed that maximum error allowable was equal to ½ distance between scale marks, unidirectional.

[4] Assumed conservative estimate that maximum error allowable was equal to the entire scale length, but that the most probable error was ½ the length of the scale (μ = .5), normally distributed, with total scale width containing $\mu \pm 3\sigma$ (σ = .165). Also assumed that individuals' point estimates will tend to bisect their region of indifference.

[5] Analogous to "Type I" and "Type II" error, that both use the same scale to estimate two different probabilities, i.e., "false positive" and "false negative."

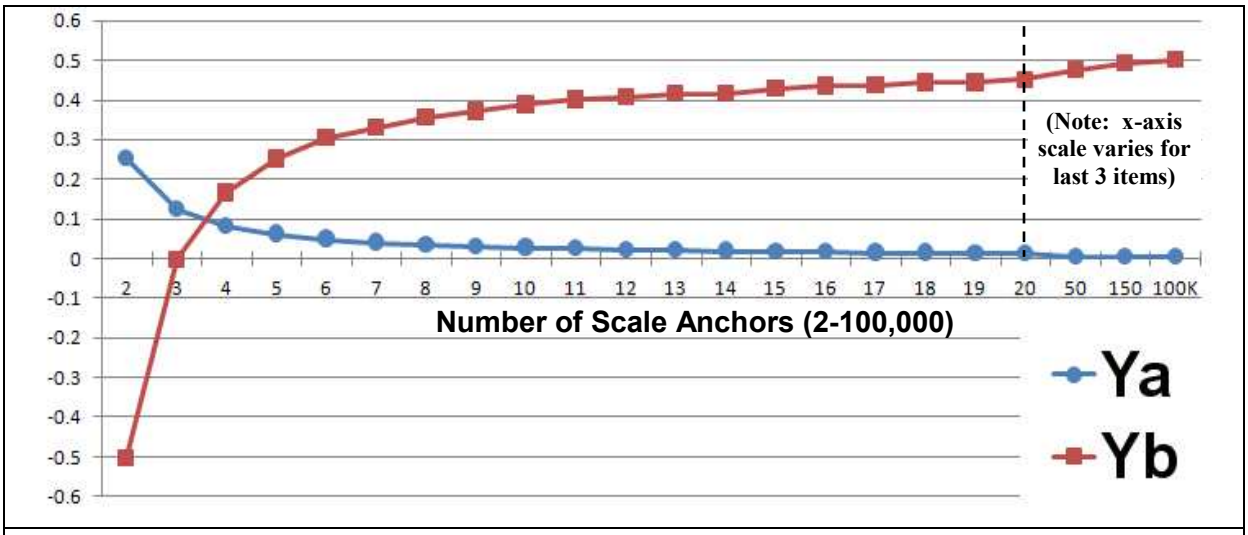

Figure 1: Simulated Average Type A and B Error Behavior as Number of Scale Anchors Increases

The results of this simulation emphasize that the choice of the number of anchors to use in a Likert-type scale results in a tradeoff between the two types of error: Type A and Type B, and that there is no optimal solution that minimizes both error types—especially considering the modern researcher's desire for interval data in order to use modern statistical programs and techniques. Clearly, another approach is desirable that might resolve this dilemma.

Study #2: Tool Development and Testing (*in progress*)

Having shown how Type A and B error rates are both related to the number of scale anchors, the next step is to begin validating functionality that might eliminate the Type A error rate, while minimizing or accounting for the Type B error rate. As outlined above, the most promising method to achieve these goals is to begin with a computer-based slider (minimizes Type A error by design), and then add functionality to it to minimize or account for Type B error rate. A literature search was unable to establish any prior empirical measurement of a region of indifference or Type B error; therefore, any use of the terms must be predicated on the establishment of their existence and behavior. The following research question guided development of a tool to measure Type B error:

- **Research Question 2:** How can a "region of indifference" (ratio of Type B error to scale size) be measured?

Creation of the Variable-Interval Slider (VIS)

The VIS is a computer-based object that a researcher might embed into an online survey. To use it, a respondent moves a mouse over or near the continuous scale, at which time a blue[6] vertical line appears on the screen. Next, the respondent drags the mouse to the location of his/her choice, clicking once with the left mouse button to anchor the blue line to the scale. At this point, if the blue line is not perfectly centered on the scale, a green region automatically appears around the blue line equal to the width between the blue line and the nearest scale anchor (see Figure 2, Item 1), as well as an equal distance in the opposite direction. Next, using either the scroll bar or the up/down keys, the respondent can resize the green region (see Figure 2, Item 2), or completely eliminate the green region to represent a point estimate, in which case the blue line turns green (see Figure 2, Item 3). If a user resizes the region past the limits allowed by the scale, then the blue line begins to slide towards the center of the scale until, when the green region is exactly equal to the scale width, the blue line is exactly centered in the scale. Using the results of this study to guide scale width in terms of the number of anchors, it appears that approximately 201 anchors might be ideal. At this point, the average estimated Type A error is .005, and a real mid-point exists. Including a real mid-point allows a given respondent to create a region of indifference that is exactly equal to the overall scale width.

[6] The color scheme using blue and green was chosen to reduce the effects of color deficiency (Bonnardel 2006).

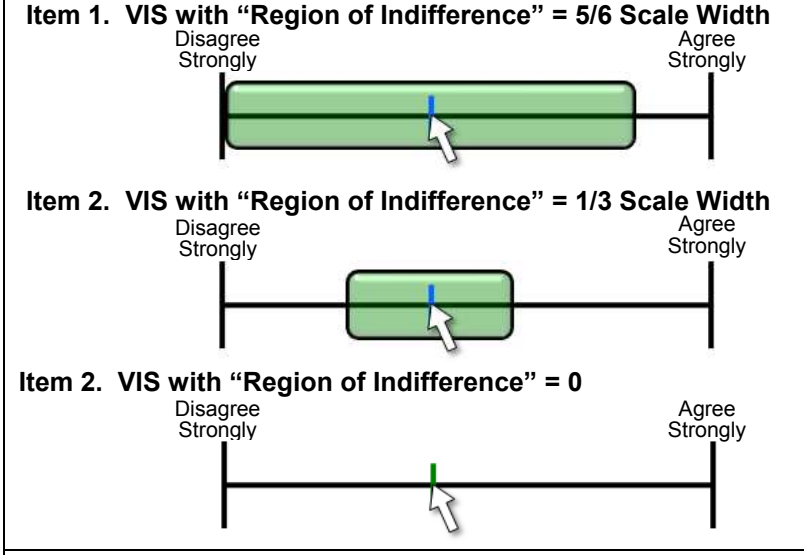

Figure 2: The VIS

Experimental Design

The primary focus of this research is to measure Type B error and compare its behavior with simulation results. Because this research intended to measure stimulus-response of an individual, a within-subjects design is used—a design that controls for between-subjects differences. A personality test is used to decrease uncertainty related to one's actual opinion (Bartholomew and Schuessler 1991), or related to an understanding of concepts surveyed, which one author noted might decrease reliability of a scale with a large number of choices (Busch 1993). The personality survey used as a vehicle for the experiment is a 44-item, short-sentence based "Big 5" personality test, originally validated with 5-anchors (John and Srivastava 2001), but used with continuous, 2-, 3-, 5-, 7-, and 20-point anchors in this study. The survey uses a unipolar scale and relatively unambiguous end-points intended to reduce the effects of respondent interpretation (Schwarz, Grayson, and Knauper 1998; Schwarz, Knauper, Hippler, Noelle-Neumann, and Clark 1991). For consistency, each of the Likert-type instruments and the VIS use only the polar anchors; eliminating intermediate anchors to prevent respondents from attaching their responses to these anchors (Mathieson and Doane 2003). Figure 3 shows an example of a 5-anchor Likert-type instrument, with dotted-outlines showing possible locations the blue selector could snap to (dotted lines do not appear in the experiment).[7]

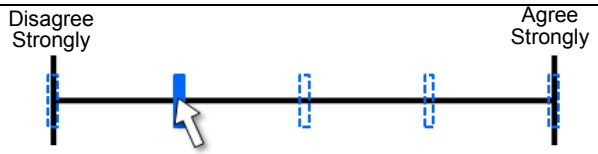

Figure 3: Example Likert-type Item Used For Comparison (5-anchor Version Shown)

The total number of items in the survey is 92, consisting of: 5 training items; 5 items testing for fatigue and drop-out; 5 items assessing demographic information and asking about disabilities; 60 experimental items; 14 non-experimental items provided as a buffer between experimental sections; 3 post-test questions to serve as a manipulation check, and to collect any qualitative information relevant to the study the subject deemed important to report. To perform the experimental manipulation, 30 of the 44 Big 5 items are asked twice, resulting in 30 pairs of questions: one Likert-type item and one VIS item. In this way, 30 pairwise comparisons are possible—enough to accommodate the 5 manipulations (VIS vs. 2-, 3-, 5-, 7-, and 20-anchor Likert-type items) with 6 questions each.

[7] Admittedly, excluding the possible anchor points from the experiment and just showing end-points could prove somewhat frustrating. For this reason, subjects will be informed that, although each scale looks exactly the same, they will not all act the same.

The Likert-type and VIS items are kept together to: a) maintain the validity of the original survey item, b) reduce the occurrence of subsequent question influence effects (Schwarz and Hippler 1995), and c) encourage the respondent to try to match the opinions to the extent possible at the instant the opinion is captured. The order of Likert-type and continuous items is alternated, with three pairs occurring in each order, reducing any effects of order on the results. The appearance of 2-, 3-, 5-, 7-, and 20-anchor Likert-type items is counterbalanced using a rectangular 5 x 6 array extension of the "Improved 5 x 5 Latin square" described by Rosenthal and Rosnow (2008, p. 541). To further reduce interactions, the remaining 14 items are interspersed between experimental items at the end of each Latin Square sequence to buffer treatments.

The experiment will be administered in a computer laboratory, with all equipment and item configurations held constant, e.g., monitors, keyboards, browser types and mouse input devices, addressing many of the confounds inherent to web-based survey development (Burkey and Kuechler 2003). Common method variance is assumed to be minimal because the independent variable is applied in the form of an experiment, even though this experiment is applied within a survey instrument. Because the effect sizes are hypothesized to be large, calculating power is straightforward.[8] Gpower (Faul, Erdfelder, Lang, and Buchner 2007), programmed with a Type I error rate of .05, and a Type II error rate of .05, indicates that the sample size required is 48.

Implications and Limitations

This research models two potentially large sources of error in survey research, one of which is not been previously identified. Considering the benefits of the tool, it might be worthwhile to further investigate a VIS or similar interface for collecting online survey data. For example, previous research showed respondents with a low socioeconomic background preferred continuous scales (Lampert 1979). Research also indicates continuous-based questions cause an individual to think more deeply about a subject (Lampert 1979). Next, continuous-based questions consistently rate high amongst participants, versus other scales (Lampert 1979; McKelvie 1978).

First, the VIS holds promise for establishing Type B error existence, as well as its behavior in individuals, question types, or populations. For methodologians and scale developers, the VIS holds promise for improving scale design, either by replacing Thurstone/Likert scales with a VIS, or by using a VIS to better validate Thurstone/Likert scales. For example, instead of using card-sorting, it might be possible to use a VIS to capture "certainty of construct," validating instrument items and assigning their representativeness of a construct based on the width or the region of indifference (with smaller regions indicating a measure that better captures the construct of interest). Because the VIS collects two data points for each question (a "center of certainty" and "region of indifference"), it might be possible to use these pieces of information to calculate measurement error based on a single question, in essence weighting a response according to how strongly the respondent feels. Perhaps the most important implication of a VIS-type design is that it holds the potential to reduce the total number of questions administered. For between-subjects designs, this could prove useful in creating better pre-tests; whereas for within-subjects designs it might allow more information to be collected, for example in a computer-based survey that branches based on previous responses, branches could be developed to activate based on an individual's level of certainty. Finally, in pre-testing Likert-type items, it might be possible to use a VIS-type design to determine the native number of scale anchors in a given population for a given question. Clearly, more research is required to evaluate these opportunities.

Considering the benefits listed above, it is also important to address remaining questions that might confound a VIS-type survey design. For example, Dillman and Smyth (2007) caution us to "use bells and whistles" sparingly in Internet survey design. Future research may wish to investigate ways to reduce any novel treatment effect an innovation like the VIS may project onto respondents. Another effect of interest, noted by some researchers, is the interplay between a respondent's choices and the end-point anchors (Dawes 2008). As this design attempts to control for these effects by choosing relatively unambiguous anchors, and by limiting an area of indifference to the confines of the scale, it is incapable of measuring this effect. The same logic holds for an attempt to measure how additional interior anchors might transform data collected using a VIS-type scale. Finally, though this study attempts to control for, and measures through self-report, any issues that may have arisen due to physical disabilities, it is important to continue to research if, and how, physical disabilities limit human-computer interaction in the online survey environment (O'Grady, Cohen, Beach, and Moody 2004; Riviere and Thakor 2005).

---

[8] Artifacts due to interaction between experimental manipulations are assumed small by comparison to the main effects, and assumed to be counterbalanced by the Latin squares-based design; therefore, they are not tested for.

References

Albaum, G., Best, R. and Hawkins, D. 1981. "Continuous vs. Discrete Semantic Differential Ratings Scales," *Psychological Reports* (49), pp. 90-97.

Alwin, D. F. 1992. "Information Transmission in the Survey Interview: Number of Response Categories and the Reliability of Attitude Measurement," *Sociological Methodology* (22), pp. 83-118.

Bartholomew, D. J., and Schuessler, K. F. 1991. "Reliability of Attitude Scores Based on a Latent Trait Model," *Sociological Methodology* (21), pp. 97-123.

Bock, D. E., Velleman, P. F., and DeVeaux, R. D. 2007. *Stats: Modeling the World* (2nd Ed.). Boston, MA: Pearson Addison-Wesley.

Bonnardel, V. 2006. "Color Naming and Categorization in Inherited Color Vision Deficiencies," *Visual Neuroscience* (23), pp. 637-643.

Burkey, J., and Kuechler, W. L. 2003. "Web-based Surveys for Corporate Information Gathering: A Bias-reducing Design Framework," *IEEE Transactions on Professional Communication* (46:2), pp. 81.

Busch, M. 1993. "Using Likert Scales in L2 Research: A Researcher Comments," *TESOL Quarterly*, (27:4), pp. 733-736.

Cox, E. P. I. 1980. "The Optimal Number of Response Alternatives for a Scale: A Review," *Journal of Marketing Research*, (17:4), pp. 407-422.

Dawes, J. 2008. "Do Data Characteristics Change According to the Number of Scale Points Used? An Experiment Using 5-point, 7-point and 10-point Scales," *International Journal of Marketing Research* (50:1), pp. 61-77.

Dillman, D. A. 2000. *Mail and Internet Surveys: The Tailored Design Method* (2nd Ed.). New York, NY: John Wiley & Sons.

Dillman, D.A., and Smyth, J.D. 2007. "Design Effects in the Transition to Web-based Surveys," *American Journal of Preventive Medicine* (32:5S), pp. S90-S96.

Dillon, W. R., Madden, T. J., and Mulani, N. 1983. "Scaling Models for Categorical Variables: An Application of Latent Structure Models," *The Journal of Consumer Research* (10:2), pp. 209-224.

Dunn, L. F. 1993. "Category Versus Continuous Survey Responses in Economic Modeling: Monte Carlo and Empirical Evidence," *The Review of Economics and Statistics* (75:1), pp. 188-193.

Faul, F., Erdfelder, E., Lang, A. G., and Buchner, A. 2007. "G*Power 3: A Flexible Statistical Power Analysis Program for the Social, Behavioral, and Biomedical Sciences," *Behavior Research Methods* (39:2), pp. 175-191.

Green, P. E., and Rao, V. R. 1970. "Rating Scales and Information Recovery: How Many Scales and Response Categories to Use?" *Journal of Marketing Research* (34:3), pp. 33-39.

House, F. N. 1934. "Measurement in Sociology" *The American Journal of Sociology* (40:1), pp. 1-11.

Jacoby, J., and Matell, M. S. 1971. "Three-point Likert Scales are Good Enough," *Journal of Marketing Research* (8:4), pp. 495-500.

John, O. P., and Srivastava, S. 2001. "The Big-Five Trait Taxonomy: History, Measurement, and Theoretical Perspectives," In *Handbook of Personality: Theory and research* (2nd ed.), Guilford, pp. 738.

Kirkpatrick, C. 1936. "Assumptions and Methods in Attitude Measurements," *American Sociological Review* (1:1), pp. 75-88.

Lampert, S. I. 1979. "The Attitude Pollimeter: A New Attitude Scaling Device," *Journal of Marketing Research* (16:4), pp. 578-582.

Lehmann, D. R., and Hulbert, J. 1972. "Are Three-point Scales Always Good Enough?" *Journal of Marketing Research*, (9:4), pp. 444-446.

Lodge, M., Cross, D. V., Tursky, B., and Tanenhaus, J. 1975. "The Psychophysical Scaling and Validation of a Political Support Scale," *American Journal of Political Science* (19:4), pp. 611-649.

Lodge, M., and Tursky, B. 1979. "Comparisons Between Category and Magnitude Scaling of Political Opinion Employing SRC/CPS Items," *The American Political Science Review* (73:1), pp. 50-66.

Mathieson, K., and Doane, D. P. 2003. "Using Fine-Grained Likert Scales in Web Surveys," *Unpublished Manuscript* (available online at: www.sba.oakland.edu/workingpapers/2003/2003-1.pdf).

McKelvie, S. J. 1978. "Graphic rating scales—How many categories?" *British Journal of Psychology* (69:2), pp. 185.

Murphy, G., and Likert, R. 1938. *Public Opinion and the Individual: A Psychological Study of Student Attitudes on Public Questions, With a Retest Five Years Later*. New York, NY: Harper & Brothers.

O'Grady, R., Cohen, C. J., Beach, G., and Moody, G. 2004. "NaviGaze: Enabling Access to Digital Media for the Profoundly Disabled," In *Proceedings of the 33rd Applied Imagery Pattern Recognition Workshop*, pp. 211-216.

Riviere, C., and Thakor, N. 1995. "Adaptive Human-machine Interface for Persons with Tremor," *Engineering in Medicine and Biology Society, IEEE 17th Annual Conference* (2), pp. 1193-1194.

Rosenthal, R., and Rosnow, R. L. 2008. *Essentials of Behavioral Research: Methods and Data Analysis* (3rd ed.). New York, NY: McGraw-Hill.

Russell, C. J., and Bobko, P. 1992. "Moderated Regression Analysis and Likert Scales Too Coarse for Comfort," *Journal of Applied Psychology* (77:3), pp. 336-342.

Schwarz, N., Grayson, C. E., and Knauper, B. 1998. "Formal Features of Rating Scales and the Interpretation of Question Meaning," *International Journal of Public Opinion Research* (10:2), pp. 177.

Schwarz, N., and Hippler, H. J. 1995. "Subsequent Questions May Influence Answers to Preceding Questions in Mail Surveys," *Public Opinion Quarterly* (59:1), pp. 93.

Schwarz, N., Knauper, B., Hippler, H. J., Noelle-Neumann, E., and Clark, L. 1991. "Rating Scales: Numeric Values May Change the Meaning of Scale Labels," *Public Opinion Quarterly* (55:4), pp. 570.

Shadish, W. R., Cook, T. D., and Campbell, D. T. 2002. *Experimental and Quasi-Experimental Designs*. Boston, MA: Houghton Mifflin.

Sherman, M. 1932. "Theories and Measurement of Attitudes," *Child Development* (13:1), pp. 15-28.

Stevens, S. S. 1946. "On the Theory of Scales of Measurement," *Science* (103:2684), pp. 677-680.

Stroud, J. B., Mollenkopf, W. G., Gerberich, J. R., Symonds, P. M., Sells, S. B., Bayley, N., et al. 1956. "Educational Measurements," *Review of Educational Research* (26:3), pp. 268-291.

Thurstone, L. L. 1928. "Attitudes Can Be Measured," *The American Journal of Sociology* (33:4), pp. 529-554.

Thurstone, L. L. 1959. "Psychophysical Analysis," In *The Measurement of Values*, L. L. Thurstone (Ed.), Chicago, IL: University of Chicago Press, pp. 19-38.

Thurstone, L. L., and Chave, E. J. 1929. *The Measurement of Attitude*. Chicago, IL: University of Chicago Press.

Townsend, J. T., and Ashby, F. G. 1984. "Measurement Scales and Statistics: The Misconception Misconceived," *Psychological Bulletin* (96:2), pp. 394-401.

van Schaik and Ling 2003. "Using On-line Surveys to Measure Three Key Constructs of the Quality of Human-computer Interaction in Web Sites: Psychometric Properties and Implications," *International Journal of Human-Computer Studies* (59), pp. 545–567.

van Schaik and Ling 2007. "Design Parameters of Rating Scales for Web Sites," *ACM Transactions on Computer-Human Interaction* (14:1), pp. 1-30.